ESCAP II: Characteristics of Census Imputations

Signe I. Wetrogan
and
Arthur R. Cresce

Population Division

USCENSUSBUREAU

Helping You Make Informed Decisions

CONTENTS

CHARACTERISTICS OF CENSUS IMPUTATIONS

EXECUTIVE SUMMARY

For Census 2000, the proportion of people for whom the basic 100-percent characteristics were imputed was higher than for the 1990 Census. This report examines the age, sex, race, and Hispanic origin characteristics assigned to the people whose 100-percent characteristics were totally imputed in Census 2000. For this analysis, the universe is restricted to the household population excluding remote Alaska.

Is the level of differential undercount affected by census imputations?

There is no evidence that imputation had an important impact on measures of differential undercount.

What were the sources of the imputations?

Approximately 5.8 million people (2.1 percent of the total population) had all their 100-percent characteristics imputed in Census 2000. They can be disaggregated further into two categories:

C   The count imputation process adds people to the census. For these imputed people, their characteristics are necessarily imputed also. In Census 2000, 1.2 million people (0.4 percent of the total population) were counted through count imputation.

C   The whole-person characteristics imputation process supplies all the characteristics data for people already counted directly through the census enumeration process. In Census 2000, 4.6 million people (1.7 percent of the total population) were imputed through the whole-person characteristics imputation process, including 2.3 million people within households with reported characteristics for some members, and 2.3 million people in households where everyone had their characteristics imputed.

There were five basic categories of census cases that required an imputation procedure to fill in missing data as to the person counts and/or all 100-percent person characteristics items. Three categories result from the count imputation process-

1) Status Imputation - no information about the housing unit;
2) Occupancy Imputation - existence of housing unit confirmed but no information as to occupancy status; and
3) Household Size Imputation - existence of occupied unit confirmed but no information as to household count.

Two categories result from the total whole-person characteristic imputation process where the

household count is known -
    1) Whole Household Imputations - all characteristics imputed for the entire household; and
    2) Within Household Imputation - all characteristics imputed for some, but not all, people in
    the household based on characteristics of other people in the household.

How were characteristics assigned to the population who had all their basic 100-percent characteristics imputed in Census 2000?

To assign characteristics to the population who had all their 100-percent data items imputed, the Census Bureau adopted a procedure which assigned characteristics based on either other household members with data-defined characteristics or on the nearest neighbor with data-defined characteristics (hot deck method). Data-defined persons are defined as records that contain two or more of the 100-percent population data items for a person or contain the person's name. Consequently, the characteristics assigned to the imputed population should be consistent with the characteristics of data-defined people.

Were the age, sex, race, and Hispanic origin characteristics assigned to the population requiring imputations similar to the age, sex, race, and Hispanic origin characteristics of the data defined population?

The age, race, and sex characteristics of the population requiring some form of imputation were similar to the characteristics of the data-defined population with the exception of the age category under age 18. The proportion of the population under age 18 in the data-defined population was 26 percent, where as in the imputed population, the proportion of the population under age 18 was 39 percent. The higher proportion occurred primarily in the within-household imputation universe.

Can the high proportion of the within household imputation universe assigned to the under age 18 category be explained?

Mail returns accounted for sixty-seven percent of the within-household imputation universe. It mostly reflected the large households (7 or more people) that were not accommodated by the 6-person mail-return questionnaire and for which the telephone followup operation was unable to supply the missing data for the additional household members. The remaining cases were from enumerator-completed questionnaires, where the enumerator was unable to collect information about all household members, or when the continuation form count could not be linked to its parent form for a large household. The characteristics for this population could be assigned from other household members and from other data-defined households. Because we most often are assigning age to the sixth and seventh person within these large households, and because children are generally listed after adults on questionnaires filled out by respondents, we would expect to assign a high proportion to the under 18 age group.

BACKGROUND

For Census 2000, the proportion of people for whom the basic 100-percent characteristics were imputed was higher than for the 1990 Census. Approximately 5.8 million people (2.1 percent of the total population) had all their 100-percent characteristics imputed in Census 2000.

Approximately 1.2 million people, or 0.4 percent of the total population, were added to the census count through a count-imputation process. People added to the count-imputation process necessarily had their person characteristics data imputed as well. The Census Bureau used count imputation for three categories of cases in Census 2000:

C Household Size Imputation - The Census Bureau imputed a population count for a housing unit when Bureau records indicated that the housing unit was occupied, but had insufficient information as to the number of individuals residing in the unit.

C Occupancy Imputation - When Census Bureau records indicated that a housing unit existed, but did not provide sufficient information to classify it definitely as either occupied or vacant, the Census Bureau imputed occupancy status (occupied or vacant). For a unit imputed as occupied, household size was also imputed.

C Status Imputation - When the Census Bureau records had insufficient information about whether an address represented a valid, non-duplicated housing unit, the Census Bureau imputed the status of the unit (occupied, vacant, or delete). For a unit imputed as occupied, household size was also imputed.

Approximately 4.6 million people, or 1.7 percent of the total population, were counted directly through the census enumeration process, but had all of their person characteristics data imputed because this information was missing from the census records. The Census Bureau imputed whole person characteristics for categories of cases in Census 2000:

C Whole Household Imputation - These households contained no data-defined people and thus required that all of the basic 100-percent characteristics data be imputed for each of the household members.

C Within Household Imputation - These households contained at least one data-defined person. However, they also contained people who were missing all of the basic 100-percent data items. The imputation process used information on other persons within the household to assign the missing values.

To assign characteristics to the population who had all their 100-percent data items imputed, the Census Bureau adopted a procedure which assigned characteristics based on either other household members with data-defined characteristics or on the nearest neighbor with data-defined characteristics (hot deck method). As such, the characteristics assigned to the imputed

population should be consistent with the characteristics of data-defined people.

This report examines the age, sex, race, and Hispanic origin characteristics assigned to the people whose 100-percent characteristics were totally imputed in Census 2000. For this analysis, the universe was restricted to the household population excluding remote Alaska. In examining the characteristics of the census imputation population, we attempted to provide insight to the question: Was the level of differential undercount affected by census imputations?

RESULTS

Census Imputations by Type

Approximately 5.8 million people had all of their 100-percent characteristics imputed. Approximately 1.2 million of these people (0.4 percent of the total population) were added to the census count through a count-imputation process (Table 1). Of this number, 496,000 were added through the household-size imputation process, 415,000 were added through the status-imputation process, and 260,000 were added through the occupancy-imputation process.

The majority of the people who had all of their 100-percent characteristics imputed (4.6 million people, or 1.7 percent of the total population) in Census 2000 were added through either the whole-household imputation process or within-household imputation process. The numbers added through the two processes were roughly equal.

Characteristics by Age

The age characteristics of the population requiring some form of imputation were similar to those of the data-defined population with the exception of the age category under age 18 (Figure 1). The proportion of the population under age 18 in the data-defined population was 26 percent. In the imputed population, the proportion of the population under age 18 was almost 39 percent.

Upon closer examination, the higher proportion of the population under age 18 in the imputed population is due to the high proportions of younger people in the within-household imputation universe. Fifty-eight percent of the within-household imputation population were assigned to the under age 18 population, compared to 26 percent in the data-defined population.

Examining the source of the within-household imputation population provides an explanation for this high proportion of the population being assigned to the under 18 age category. Sixty-seven percent of the within-household imputation universe was from mail returns. These mostly reflect the large households (7 or more people) that were not accommodated by the 6-person mail-return questionnaire and for which the telephone followup operation was unable to supply the missing data. The remaining cases were from enumerator-completed questionnaires where the enumerator was unable to collect information about all household members, or the continuation form was unable to be linked to its parent form for some large households.

2

The characteristics for this population could be assigned from other household members and from other data-defined households. Because we most often are assigning age to the sixth and seventh persons in these large households, we would expect to assign a high proportion to the under 18 age group.

Characteristics by Race

The race characteristics of the various populations requiring some form of imputation are similar to those of the data-defined population (Figure 2). Nonblacks accounted for 88 percent of the data-defined population. This compares to 81 percent of the within-household imputation population and 77 percent of the whole-household imputation population. Among the population requiring some form of count imputation, 76 percent of the household-size imputation population was Nonblack compared to 85 percent of the occupancy-imputation population and 89 percent of the status-imputation population.

Characteristics by Hispanic Origin

The Hispanic origin characteristics of the various populations requiring some form of imputation are generally similar to those of the data-defined population (Figure 3). The only exception is the within-household imputation population. The proportion of the within-household population that was Hispanic origin was 33 percent compared to 12 percent of the data-defined population.

The source of the within-household imputation population provides a likely explanation for the higher proportion identified as Hispanic origin. As discussed above, many of these cases reflect the large households (7 or more people) that were not accommodated by the 6-person mail-return questionnaire and for which the telephone followup operation was unable to supply the missing data. The characteristics of this population could be assigned from other household members and from other data-defined households.

CONCLUSION

Was the level of differential undercount affected by census imputations?

For most of the census-imputation population, the age, sex, race, and Hispanic origin characteristics are similar to those of the data-defined population. While we did identify some outliers among the within-household imputation population, the source of this population provides a plausible explanation for the higher than expected proportion under age 18 and the higher than expected proportion of Hispanic origin. Thus, we can find no evidence that the imputation process had an important impact on measures of differential undercount.
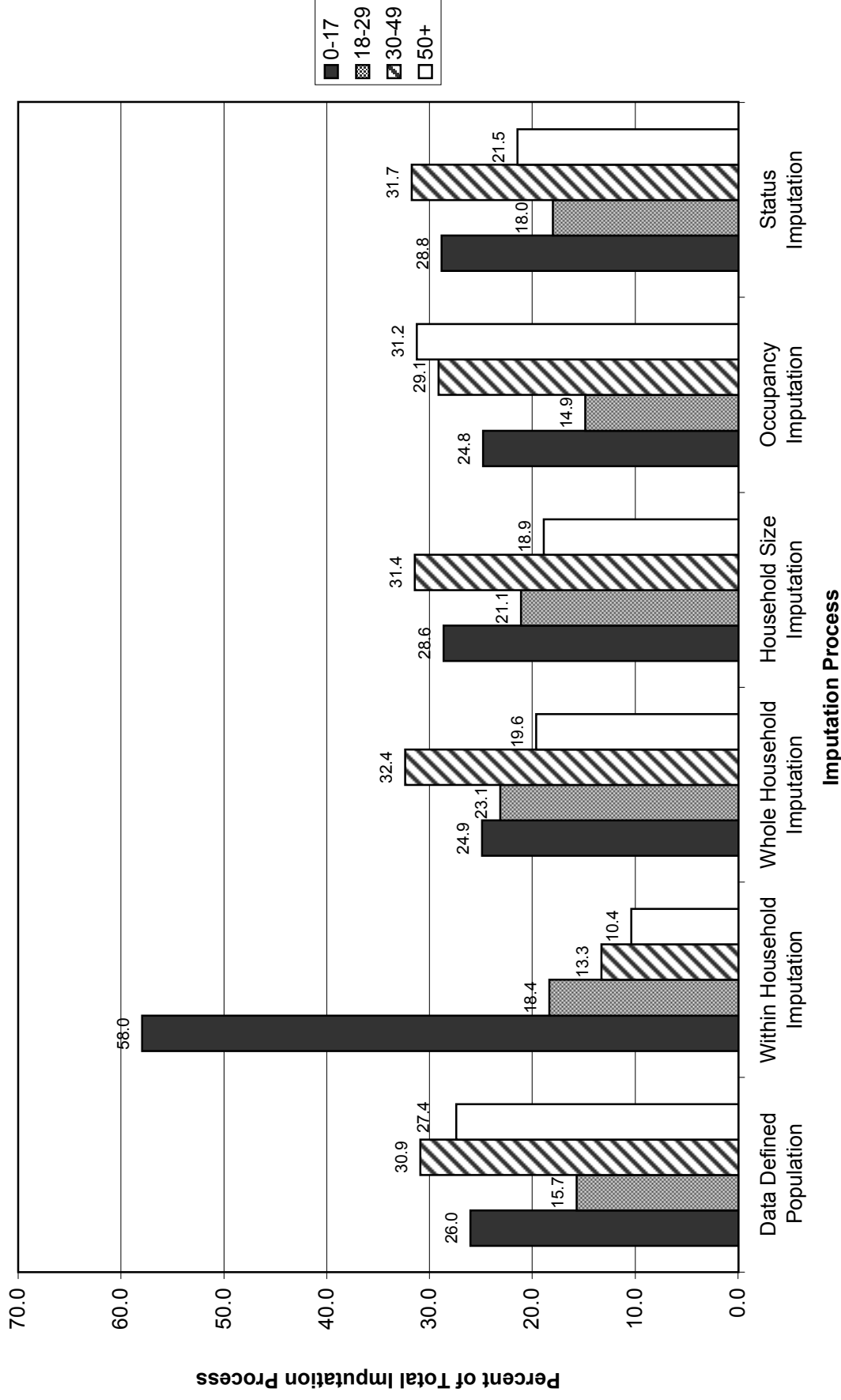
Figure 1

Census 2000 Imputations by Age
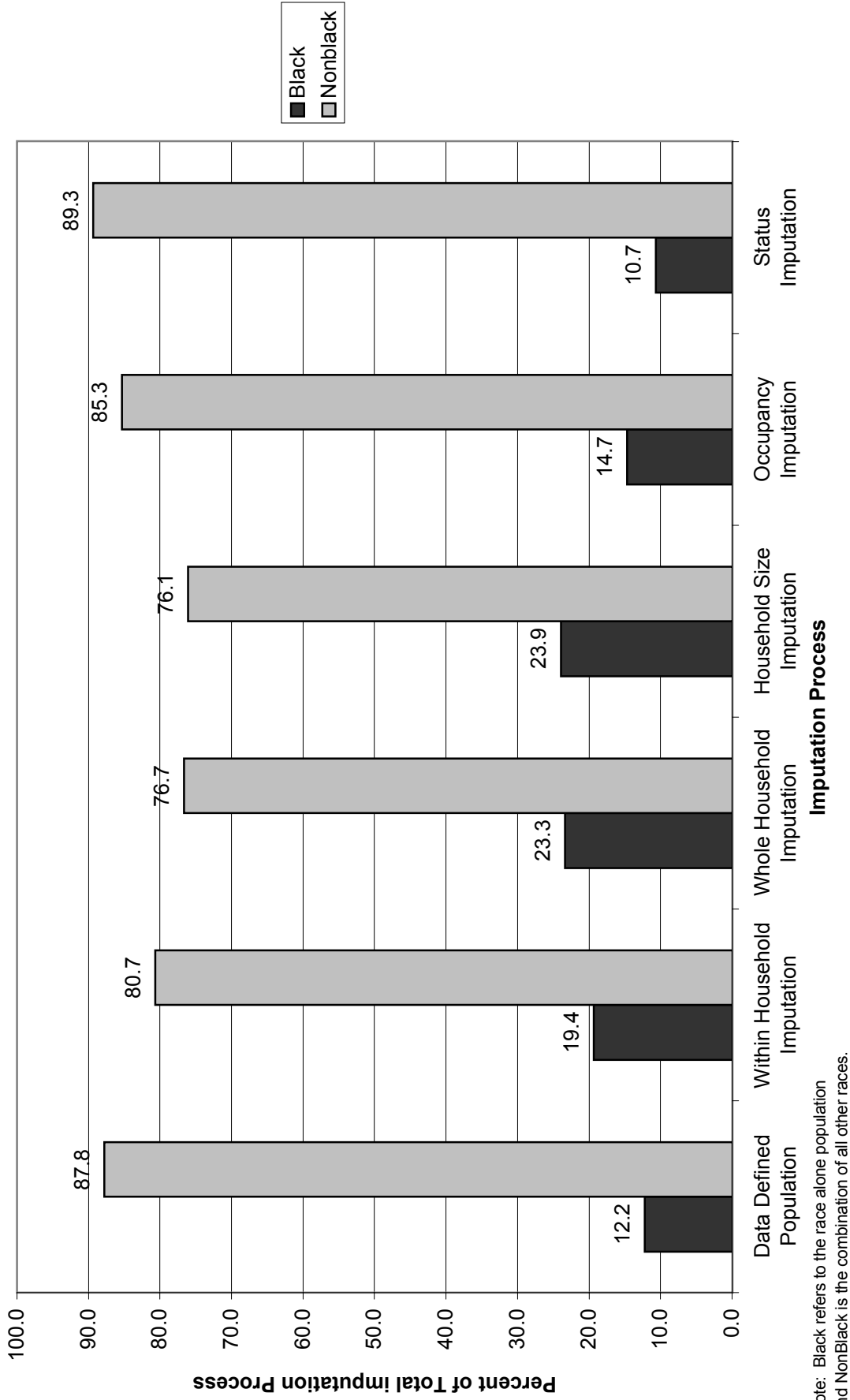
Figure 2

Census 2000 Imputations by Race

Note: Black refers to the race alone population
and NonBlack is the combination of all other races.

Figure 3

Census 2000 Imputations by Hispanic Origin
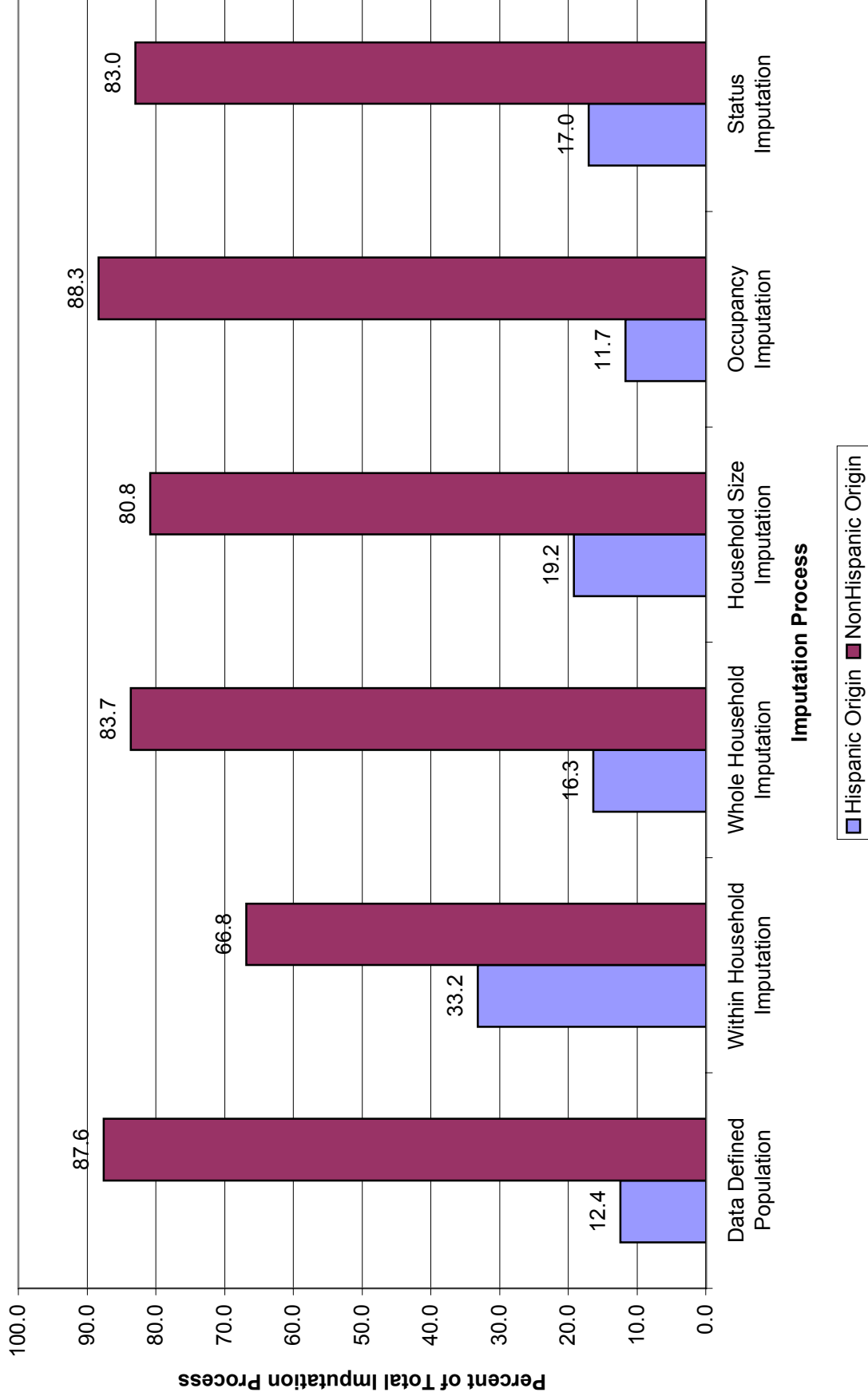
Table 1: Census 2000 Imputations by Type

Subject	Total population	Data-defined population	Imputed Populaton					
			Whole-person characteristics			Count imputation		
			Total	Within household imputations	Whole household imputation	Household size imputation	Occupancy imputation	Status imputation
Population	##########	##########	5,766,893	2,326,761	2,268,946	495,553	260,420	415,213
Percent of tot	100.0	97.9	2.1	0.9	0.8	0.2	0.1	0.2

Note: Household population excluding remote Alaska.